

# Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods

Ronald D. Fricker, Jr.\*  
and  
David Banschbach†

June 1, 2009

## Abstract

We describe a methodology for optimizing a threshold detection-based biosurveillance system. The goal is to maximize the system-wide probability of detecting an “event of interest” against a noisy background, subject to a constraint on the expected number of false signals. We use non-linear programming to appropriately set detection thresholds taking into account the probability of an event of interest occurring somewhere in the coverage area. Using this approach, public health officials can “tune” their biosurveillance systems to optimally detect various threats, thereby allowing practitioners to focus their public health surveillance activities. Given some distributional assumptions, we derive a 1-dimensional optimization methodology that allows for the efficient optimization of very large systems. We demonstrate that optimizing a syndromic surveillance system can improve its performance by 20-40 percent.

KEYWORDS: Biosurveillance, syndromic surveillance, bioterrorism, public health, optimization, Shewhart chart.

## 1 Introduction

*Biosurveillance* is the practice of monitoring populations – human, animal, and plant – for the outbreak of disease. Often making use of existing health-related data, one of the principle objectives of biosurveillance systems has been to give early warning of bioterrorist attacks or other emerging health conditions (CDC 2004). The Centers for Disease Control and Prevention (CDC) as well as many state and local health departments around the United States are developing and fielding *syndromic surveillance* systems, one type of biosurveillance.

---

\*Naval Postgraduate School, Operations Research Department, Monterey, CA

†Naval Postgraduate School, Operations Research Department, Monterey, CA

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>01 JUN 2009</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2009 to 00-00-2009</b>	
4. TITLE AND SUBTITLE <b>Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Naval Postgraduate School, Operations Research Department, Monterey, CA, 93943</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <b>see report</b>					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>29</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

A *syndrome* is “A set of symptoms or conditions that occur together and suggest the presence of a certain disease or an increased chance of developing the disease” (International Foundation For Functional Gastrointestinal Disorders 2006). In the context of syndromic surveillance, a syndrome is a set of non-specific pre-diagnosis medical and other information that may indicate the health effects of a bioterrorism agent release or natural disease outbreak. See, for example, Syndrome Definitions for Diseases Associated with Critical Bioterrorism-associated Agents (CDC 2003). The data in syndromic surveillance systems may be clinically well-defined and linked to specific types of outbreaks, such as groupings of ICD-9 codes from emergency room “chief complaint” data, or only vaguely defined and perhaps only weakly linked to specific types of outbreaks, such as over-the-counter sales of cough and cold medication or absenteeism rates.

Since its inception, one focus of syndromic surveillance has been on early event detection: gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak. Such early event detection is not supposed to provide a definitive determination that an outbreak is occurring. Rather, it is supposed to signal that an outbreak *may* be occurring, indicating a need for further evidence or triggering an investigation by public health officials (i.e., the CDC or a local or state public health department). See Fricker (2008), Fricker (2007), and Fricker & Rolka (2006) for more detailed exposition and discussion.

BioSense and EARS are two biosurveillance applications currently in use. The first is a true system, in the sense that it is comprised of dedicated computer hardware and software that collect and evaluate data routinely submitted from hospitals. The second is a set of software programs that are available for implementation by any public health organization.

- *BioSense* was developed and is operated by the National Center for Public Health Informatics of the CDC. It is intended to be a United States-wide electronic biosurveillance system. Begun in 2003, BioSense initially used Department of Defense and Department of Veterans Affairs outpatient data along with medical laboratory test results from a nationwide commercial laboratory. In 2006, BioSense began incorporating data from civilian hospitals as well. The primary objective of BioSense is to “expedite event recognition and response coordination among federal, state, and local public health and health care organizations” (Fricker 2008, CDC 2006a, Tokars 2006a,b). As of May 2008, BioSense was receiving data from 563 facilities (CDC 2008).
- *EARS* is an acronym for Early Aberration Reporting System. Developed by the CDC, EARS was designed to monitor for bioterrorism during large-scale events that often have little or no

baseline data (i.e., as a short-term drop-in surveillance method) (CDC 2006b). For example, the EARS system was used in the aftermath of Hurricane Katrina to monitor communicable diseases in Louisiana, for syndromic surveillance at the 2001 Super Bowl and World Series, as well as at the Democratic National Convention in 2000 (Toprani *et al.* 2006, Hutwagner *et al.* 2003). Though developed as a drop-in surveillance method, EARS is now being used on an on-going basis in many syndromic surveillance systems.

A characteristic of many syndromic surveillance systems is that the data collection locations (typically hospitals and clinics) are in fixed locations that may or may not correspond to a particular threat of either natural disease or bioterrorism. Thus, in order to guard against any eventuality, syndromic surveillance system designers and operators are inclined to enlist as many hospitals and clinics, and to solicit as much data from them, as possible. However, as the sources and types of data being monitored proliferate in a biosurveillance system, then so do the false positive signals from the systems. Indeed, false positives have become an epidemic problem for some systems. As one researcher (Shmueli 2006) said, “...most health monitors...learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems – there is nearly an alarm every day!”

Our research provides a methodology which, if implemented, would allow public health officials “tune” their biosurveillance systems to optimally detect various threats while explicitly accounting for organizational resource constraints available for investigating and adjudicating signals. This allows practitioners to focus their public health surveillance activities on locations or diseases that pose the greatest threat at a particular point in time. Then, as the threat changes, using the same hospitals and clinics, the system can subsequently be tuned to optimally detect other threats. With this approach large biosurveillance systems are an asset.

The methodology assumes spatial independence of the data and temporal independence of the signals. The former is achieved by monitoring the residuals from some sort of model to account for and remove the systematic effects present in biosurveillance data. The assumption is that, while it is likely that raw biosurveillance data will have spatial correlation, once the systematic components of the data are removed the residuals will be independent. The latter is achieved by employing detection algorithms that only depend on data from the current time period.

It is worth emphasizing that our focus is on how to optimally set threshold levels for detection in an *existing* system, rather than how to design a new system. This is something of a unique problem for syndromic surveillance systems, meaning that in many other types of sensor systems,

one might design a system for a specific, unchanging threat or change the location of the sensors to respond to a changing threat. But in syndromic surveillance systems, where we can think of each hospital or clinic as a fixed biosurveillance “sensor” for a particular location or population, the sensor locations cannot be changed. Part of the solution is to adjust the way the data from the sensors are monitored.

## 1.1 Threshold Detection Methods

In this work, we define a *threshold detection* method as an algorithm that generates a binary output, signal or no signal, given that some function of the input or inputs exceed a pre-defined threshold level. In addition, for the methods we consider, inputs come in discrete time periods and the decision to signal or not is based only on the most recent input or inputs. That is, the methods do not use historical information in their signal determination; they only use the information obtained at the current time period.

In the quality control literature, the *Shewhart chart* is such a threshold detection method. At each time period a measurement is taken and plotted on a chart. If the measurement exceeds a pre-defined threshold a signal is generated. However, if the measurement does not exceed the threshold then the process is repeated at the next time period, and continues to be repeated until such time as the threshold is exceeded. See Shewhart (1931) or Montgomery (2001) for additional detail. A sonar detection algorithm based on signal excess is also an example of threshold detection. See Washburn (2002) and references therein for a discussion.

Threshold detection methods are subject to errors, either signalling that an event of interest occurred when it did not, or failing to signal when in fact the event of interest did occur. In classical hypothesis testing, these errors are referred to as *Type I* and *Type II* errors, respectively. A Type I error is a false signal and a Type II error is a missed detection. In threshold detection, setting the threshold requires making a trade-off between the probability of false signals and the probability of a missed detection. A *receiver operating characteristic* (or ROC) curve is a plot of the probability of false signal versus probability of detection (one minus the probability of a missed detection) for all possible threshold levels. See Washburn (2002, Chpt. 10) and the references therein for additional discussion.

## 1.2 Optimizing Sensor Systems

Optimizing a system of threshold detection-based sensors, in the sense of maximizing the probability of detecting an event of interest somewhere in the region being monitored by the system, subject to a constraint on the expected number of system-wide false signals, has not been done. Washburn (2002, Chpt. 10.4) introduces the idea of optimizing the threshold for a single sensor, parameterizing the problem in terms of the cost of a missed detection and the cost of a false signal, and seeks to minimize the average cost “per look.” He concludes that “In practice, the consequences of the two types of error are typically so disparate that it is difficult to measure  $c_1$  [cost of a missed detection] and  $c_2$  [cost of a false signal] on a common scale. For this reason, the false alarm probability is typically not formally optimized in practice.”

Kress *et al.* (2008) develop a methodology for optimizing the employment of non-reactive arial sensors. In their problem the goal is to optimize a mobile sensor’s search path in order to identify the location or locations of fixed targets with high probability. By dividing the search region into a grid of cells, Kress *et al.* use a Bayesian updating methodology combined with an optimization model that seeks to maximize the probability of target location subject to a constraint on the number of looks by the sensors. Their work differs from ours in a number of important respects, including that their sensors can have multiple looks for a target, there may be multiple targets present, and the use of Bayesian updating to calculate the probability of a target being present in a particular grid cell. In contrast, in our problem the sensors are fixed, they can only take one look per period, and at most one “event of interest” can occur in any time period.

One active area of research is how to combine threshold rules for systems of sensors in order to achieve high detection rates and low false positive rates compared to the rates for individual sensors. For example, Zhu *et al.* (2007) consider a system of threshold detection sensors for which they propose a centralized “threshold-OR fusion rule” for combining the individual sensor node decisions. In this work Zhu *et al.* (2007) allow that multiple sensors may detect the presence of the target with signals of varying strength and their objective is to combine the decisions made by individual sensors to achieve system detection performance beyond a weighted average of individual sensors. Their work builds upon the research of Chair & Varshney (1986) who, via a log-likelihood ratio test, derived a fusion rule that combines the decisions from the  $n$  individual threshold detection sensors while minimizing the overall probability of error.

### 1.3 Paper Organization

The paper is organized as follows. In Section 2 we formulate the general problem and its solution via an  $n$ -variable nonlinear program, illustrate the methodology on some simple examples, and then derive an equivalent 1-dimensional optimization problem given some distributional assumptions. In Section 3 we apply the methodology to our motivating problem, biosurveillance, using some hypothetical syndromic surveillance systems. And, in Section 4 we summarize and discuss our results, including directions for future research.

## 2 Problem Formulation

Consider a system of  $n$  sensors and let  $X_{it}$  denote the output from sensor  $i$ ,  $i = 1, \dots, n$ , at time  $t$ ,  $t = 1, 2, \dots$ . Sensor outputs occur at discrete time periods and each sensor has exactly one output per time period.

Assume that when no event of interest is present anywhere in the system the  $X_{it}$  are independent and identically distributed,  $X_{it} \sim F_0$  for all  $i$  and all  $t$ . If an event of interest occurs at time  $\tau$ , then  $X_{i\tau} \sim F_1$  for exactly one  $i$ . A signal is generated at time  $\tau^*$  when  $X_{i\tau^*} \geq h_i$  for one or more  $i$ , where the thresholds  $h_i$  can be set separately for each sensor.

Further assume that there is some distribution on the probability that an event of interest will occur at sensor  $i$ 's location,  $p_i$ ,  $\mathbf{p} = \{p_1, p_2, \dots, p_n\}$ , where  $\sum_i p_i = 1$ . Note that  $\mathbf{p}$  is a conditional probability: it is the probability an event occurs in sensor  $i$ 's location given that an event occurs somewhere in the system.

The goal is to choose thresholds that maximize the probability of detecting the event of interest, given one occurs somewhere in the region according to  $\mathbf{p}$ , subject to a constraint on the conditional expected number of system-wide false signals per time period.

For sensor  $i$  at time  $t$ , the probability of a true signal is

$$\mathbb{P}(\text{signal} | \text{event of interest occurs at sensor } i\text{'s location}) = \int_{z=h_i}^{\infty} f_1(z) dz = 1 - F_1(h_i) = \delta_i, \quad (1)$$

and the probability of a false signal at sensor  $i$  is

$$\mathbb{P}(\text{signal} | \text{no event of interest at sensor } i\text{'s location}) = \int_{z=h_i}^{\infty} f_0(z) dz = 1 - F_0(h_i) = \alpha_i. \quad (2)$$

Thus, given that an event occurs in a particular time period, the probability the system detects

the event is  $\sum_{i=1}^n \delta_i p_i$ . Further, given that no event occurs, the expected number of false signals in a particular time period is  $\sum_{i=1}^n \alpha_i$ .

This latter quantity deserves further explanation. The  $\sum_{i=1}^n \alpha_i$  is the expected number of false signals *given that no event occurs anywhere in the system*. As such, it is a measure of the cost of operating the system for an event-free time period.

Define  $\mathbf{h} = \{h_1, \dots, h_n\}$ . Then we can pose the problem as the following nonlinear program (NLP),

$$\begin{aligned} \max_{\mathbf{h}} \quad & \sum_{i=1}^n [1 - F_1(h_i)] p_i \\ \text{s.t.} \quad & \sum_{i=1}^n [1 - F_0(h_i)] \leq \kappa, \end{aligned} \tag{3}$$

where  $\kappa$  is the limit on the average number of false signals per period of time. We will use the shorthand notation  $P_d(\mathbf{h})$  for the objective function, sometimes suppressing the dependency on the vector of thresholds  $\mathbf{h}$ .

Note that in this formulation of the problem we are maximizing the probability of detecting a single event that occurs somewhere in the system. This is a conservative detection probability, in the sense that if multiple events occur simultaneously, or if a single event is so large that it is detected by multiple sensors, then the actual probability of detection will be greater than  $P_d(\mathbf{h})$ .

Also note that within the NLP formulation, additional constraints can be added, depending on the requirements of the particular system or problem. For example, a constraint specifying a lower bound on the conditional probability of detection for sensor  $i$ ,  $\delta'_i$ , in the form of an upper bound on the threshold for sensor  $i$ , could be added:  $h_i \leq F_1^{-1}(1 - \delta'_i)$ . Or a constraint specifying an upper bound on the probability of a false signal for sensor  $i$ ,  $\alpha'_i$ , in the form of a lower bound on the threshold for sensor  $i$ , could be added:  $h_i \geq F_0^{-1}(1 - \alpha'_i)$ .

## 2.1 The Biosurveillance Problem

Consider a biosurveillance system of  $n$  hospitals, each located in a separate geographic region, and each feeding data on a particular syndrome into a syndromic surveillance system. Within the syndromic surveillance system each stream of data from each hospital is monitored with a Shewhart chart. Hence, we can think of each hospital-Shewhart chart combination as a biosurveillance threshold detection-based “sensor.”

Syndromic surveillance data is generally autocorrelated, with various trends and other systematic components that correspond to day-of-the-week, seasonal, and other effects. We assume that such



systematic components of the data can be appropriately modeled and thus accounted for and removed from the data. See, for example, Fricker *et al.* (2008a) and Fricker *et al.* (2008b) where adaptive regression was used to remove the systematic effects from syndromic surveillance data. We then assume that the Shewhart charts are used to monitor the standardized residuals from such a model and that the residuals can be assumed to be independently distributed according to a standard normal distribution. Finally, we assume that a disease outbreak will manifest as a step increase in the mean of the residual distribution.

Thus, based on these assumptions, we have that:

- There are  $n$  independent “sensors,” each corresponding to a hospital in a separate geographic region, each using a threshold detection algorithm (Shewhart chart) to monitor for a disease outbreak or bioterrorism attack.
- An attack in any region will manifest itself in the same way at each hospital, at least in terms of the standardized residuals being monitored. So  $X_i \sim F_0 = N(0, 1)$  when there is no bioterrorism attack and  $X_j \sim F_1 = N(\gamma, 1)$  when an attack occurs in the region served by hospital  $j$ .
- Therefore, for sensor  $i$  with threshold  $h_i$  the probability of a false signal is

$$\mathbb{P}(\text{signal} | \text{no attack in region } i) = \int_{x=h_i}^{\infty} f_0(x) dx = 1 - \Phi(h_i),$$

where  $\Phi(h_i)$  denotes the cdf for the standard normal evaluated at  $h_i$ , and the probability of a true signal is

$$\mathbb{P}(\text{signal} | \text{attack in region } i) = \int_{x=h_i}^{\infty} f_1(x) dx = \int_{x=h_i-\gamma}^{\infty} f_0(x) dx = 1 - \Phi(h_i - \gamma).$$

So, given the above assumptions, the general NLP of Equation (3) can be expressed as

$$\begin{aligned} \min_{\mathbf{h}} \quad & \sum_{i=1}^n \Phi(h_i - \gamma) p_i \\ \text{s.t.} \quad & \sum_{i=1}^n \Phi(h_i) > n - \kappa, \end{aligned} \tag{4}$$

where  $p_i$  is the probability of attack in region  $i$  (which we have yet to specify).

## 2.2 Optimizing Thresholds

Given an appropriate choice of  $\kappa$  in (3) or (4), the relevant question is how to set the various thresholds,  $h_1, \dots, h_n$ . In general there is no simple analytical solution, since it depends on  $F_0$  and

$F_1$ . For example, consider a system of just two sensors in which the event of interest is equally likely to occur at either sensor's location. In such a case, one might assume that the strategy that maximizes the probability of detecting the event is the one that sets equal thresholds on the two sensors. Yet, this is not necessarily so.

To illustrate, for this simple system we have  $\mathbf{p} = \{1/2, 1/2\}$  and, if we set the thresholds equally so that  $h_1 = h_2 = h$ ,

$$P_d = \sum_{i=1}^2 \frac{1}{2} [1 - F_i(h)] = 1 - F_1(h).$$

Assuming the maximum probability of detection occurs on the constraint boundary (so that the constraint can be expressed as an equality), we also have

$$\sum_{i=1}^2 \alpha_i = \sum_{i=1}^2 \kappa/2 = \kappa.$$

Now, choose some  $\epsilon$ ,  $0 < \epsilon < \kappa/2$ , and define  $\alpha'_1 = \kappa/2 - \epsilon$  and  $\alpha'_2 = \kappa/2 + \epsilon$ , so that  $\alpha'_1 + \alpha'_2 = \kappa$  still. Then, assuming  $F_0$  is continuous,  $h'_2 = F_0^{-1}(1 - \alpha'_2) > h > h'_1 = F_0^{-1}(1 - \alpha'_1)$  and

$$P'_d = \sum_{i=1}^2 \frac{1}{2} [1 - F_i(h'_i)] = 1 - \frac{F_1(h'_1) + F_1(h'_2)}{2}.$$

The result is that whether  $P_d > P'_d$ ,  $P_d = P'_d$ , or  $P_d < P'_d$  depends on the shapes of the distribution functions  $F_0$  and  $F_1$  between  $h'_1$  and  $h'_2$ . In particular, if  $F_1$  is convex between  $h'_1$  and  $h'_2$  then  $P_d > P'_d$ , and conversely, if  $F_1$  is concave between  $h'_1$  and  $h'_2$  then  $P_d < P'_d$ .

The point is that it is not obvious how one should best choose the thresholds, even in such a simple case as this with only two sensors and equal probability of attack at each sensor.

### 2.2.1 Some Illustrative Examples

Again, consider a system with only two sensors so that we can graph the objective function and the feasible region. For example, Figure 1 shows the plot of an objective function for a two-sensor system with  $F_0 = N(0, 1)$ ,  $F_1 = N(1, 1)$  and  $\mathbf{p} = \{1/2, 1/2\}$ . We can observe a number of features of the objective function for this simple problem.

First, it is clear that the function is increasing as either  $h_1$  or  $h_2$  (or both) decrease. Thus, without the constraint, the optimal solution is simply to set  $h_1 = h_2 = -\infty$ . Of course, in practice these are useless thresholds since at such settings the sensors will signal at every time period.

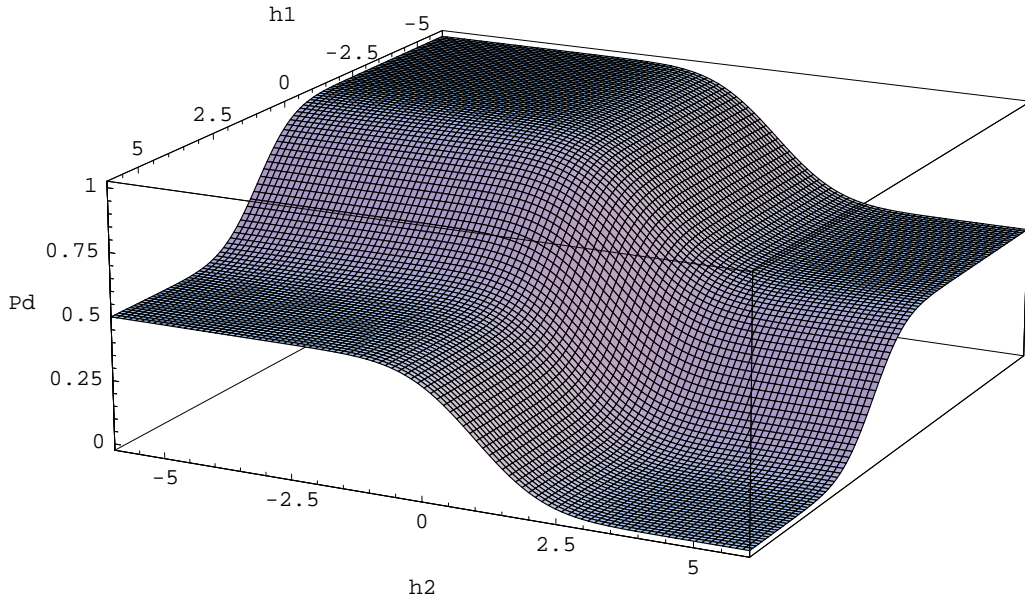


Figure 1: Plot of an objective function for  $n = 2$  with  $F_0 = N(0,1)$ ,  $F_1 = N(1,1)$  and  $\mathbf{p} = \{1/2, 1/2\}$ .

Second, there are relatively flat regions of the objective function corresponding to the tails of the  $F_1$  distribution. In these regions the objective function will be relatively insensitive to changes in the thresholds. This suggests that additional constraints can be included in the NLP restricting the thresholds to be within some reasonable domain of  $F_1$  that contains most of the dynamic range of the cumulative probability distribution. Such constraints may be useful for bounding the problem in order to facilitate convergence in an optimization package.

Figure 2 shows a view of the feasible region of the objective function for the constraint  $\alpha_1 + \alpha_2 \leq 0.1$ , where the vertical curved plane shows the boundary where  $\alpha_1 + \alpha_2 = 0.1$ . Looking at the intersection of the objective function and the vertical plane, it is visually clear that an optimal solution exists. In fact, the objective function is maximized at  $h_1 = h_2 = 1.645$ . As we will see in the next subsection, it is not an accident that the optimal solution occurs on the boundary of the feasible region.

Now consider a system of 10 hospitals, as depicted in Table 1. In this system, the event of interest is much more likely to occur at one hospital's location (hospital 1). In fact  $p_1$  is an order of magnitude greater than the probability at the next most likely hospital's location. Assuming  $F_0 = N(0,1)$  and  $F_1 = N(1,1)$ , the column labeled "Common Threshold #1" shows that the system would achieve a probability of detection of  $P_d = 0.117$  and an expected false signal rate of 0.143 signals per period

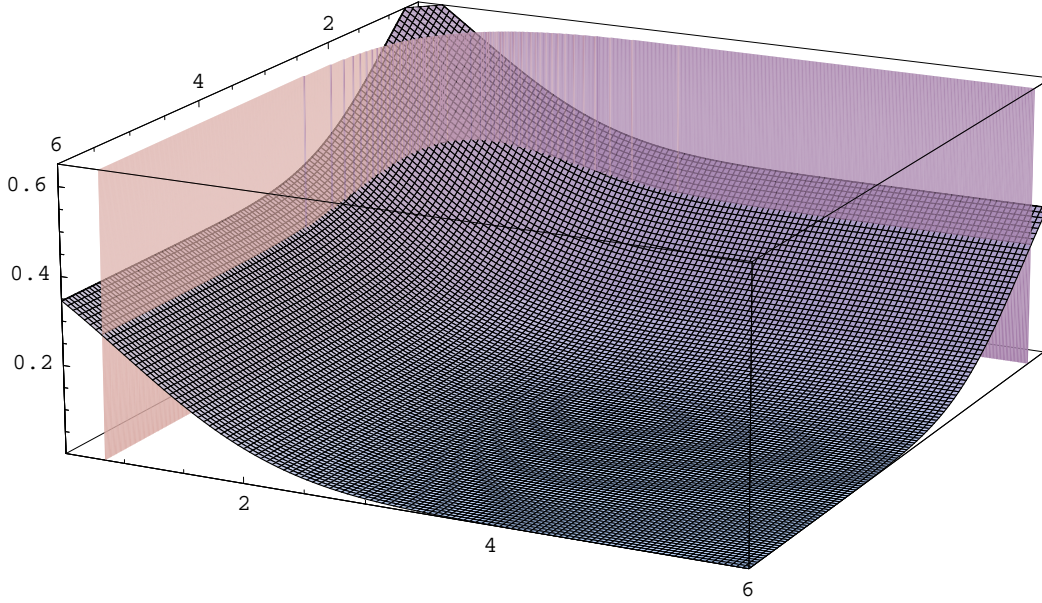


Figure 2: Plot showing the feasible region of the objective function in Figure 1, where the vertical curved plane is the boundary of the constraint  $\alpha_1 + \alpha_2 \leq 0.1$ . (The feasible region is in the foreground.)

using a common threshold of 2.189 for all hospitals. However, by optimizing the thresholds, the “Optimal Threshold” column shows that a probability of detection of  $P_d = 0.378$  can be achieved for the same expected false signal rate. This is achieved by lowering the thresholds (equivalently, increasing the probability of detecting an attack should one occur) in those locations more likely to experience an event of interest while raising the thresholds in those locations less likely to have an event of interest. Finally, the column labeled “Common Threshold #2” shows that to achieve the same  $P_d = 0.378$  with a common threshold the system would produce an expected number of false signals of almost one per period.

For a small system, with  $F_0$  and  $F_1$  normal distribution functions, it is a simple matter to express the NLP in an Excel spreadsheet using the NORMDIST function and subsequently solve it using the Solver. For this example, we used the Solver in Excel 2003 to find the optimal thresholds, which ran quickly (less than a few seconds) and reliably found the optimal solution. (Within the Solver, we used the Newton search method with Precision=  $1 \times 10^{-7}$ , Tolerance=  $5 \times 10^{-6}$ , and Convergence=  $1 \times 10^{-7}$ .) We verified the Excel solutions in Mathematica 5.0 (using the NMaximize function) and also in GAMS using the MINOS solver.

Table 1: An illustrative ten-hospital system with a specific  $\mathbf{p}$  vector. The “Optimal Threshold” column shows that  $P_d = 0.378$  can be achieved with a constraint on the expected number of false signals of one per every seven periods. The other two columns show the common thresholds that either matches the expected number of false signals at the cost of a lower  $P_d$  or that achieves the optimal  $P_d$  at the expense of increased false signals.

Hospital $i$	$p_i$	Common Threshold #1	Optimal Threshold ( $h_i$ )	Common Threshold #2
1	0.797	2.189	1.068	1.310
2	0.064	2.189	3.602	1.310
3	0.056	2.189	3.732	1.310
4	0.048	2.189	3.915	1.310
5	0.013	2.189	4.656	1.310
6	0.006	2.189	4.736	1.310
7	0.006	2.189	4.736	1.310
8	0.005	2.189	4.755	1.310
9	0.003	2.189	4.773	1.310
10	0.002	2.189	4.791	1.310
	$P_d$	0.117	0.378	0.378
	$\sum \alpha_i$	0.143	0.143	0.951

However, it is important to note that the Solver is limited to 200 adjustable cells (<http://support.microsoft.com/kb/75714>), which puts an upper bound on the number of hospitals (generically, sensors) that can be optimized using this approach. For larger systems one might consider the Excel Premium Solver, which can be used for up to 500 adjustable cells ([www.solver.com/xlsplatform.htm](http://www.solver.com/xlsplatform.htm)), but in a test with 400 hospitals the Premium Solver did not find an optimal solution after 12 hours of run-time on a fast PC. Mathematica had an even more difficult time, failing to converge on smaller systems.

The fundamental problem is that every additional sensor adds a variable to the NLP. As the dimensionality of the problem grows, more specialized optimization software such as the MINOS solver in GAMS may suffice, though very large systems will likely exceed the capacity of even these programs to solve via brute force. This suggests a need for an alternative solution methodology that reduces the dimensionality of the problem.

### 2.2.2 Reducing the Dimensionality of the Problem

Even though it is easy to show that under some relatively mild conditions the objective function in (3) is strongly quasiconvex over the constraint region, because this is a maximization problem a globally-optimal solution is not guaranteed. However, we can derive some useful theoretical properties of the constraint, particularly that the solution lies on the boundary of the constraint. Then, using this fact, and further assuming some distributional properties for  $F_0$  and  $F_1$ , we can simplify this from an  $n$ -variable optimization problem to a 1-variable optimization problem with a guaranteed optimal solution.

We begin with a simple lemma that specifies when the NLP is unconstrained.

**Lemma 1** *The NLP is unconstrained if  $\kappa \geq n$ .*

*Proof.* We first note that  $\alpha_i$  is simply the probability of a Type I error (i.e., a false signal) for sensor  $i$ . Thus, the constraint in (3) can be re-written as

$$\sum_{i=1}^n F_0(h_i) > n - \kappa.$$

Since  $0 \leq F_0(h_i) \leq 1$  for all  $h_i \in \mathbb{R}$ , the above inequality must be trivially true whenever  $\kappa \geq n$ .  $\square$

What Lemma 1 says, unsurprisingly, is that the expected number of false signals must be less than the number of sensors for the constraint to be relevant. If  $\kappa \geq n$  the maximization problem then becomes trivial: set  $h_i = -\infty$  for all  $i$  and  $P_d = \sum_{i=1}^n p_i = 1$ . In practice what this means is that each sensor produces an signal at every period and thus one is guaranteed to “detect” the event of interest. This is equivalent to a 100 percent inspection scheme in which the sensors are irrelevant.

Of course, in an actual biosurveillance system application, the constraint on the expected number of false signals will of necessity be much smaller than the number of hospitals. Signals consume resources as they must be investigated to determine whether an event of interest actually occurred, and a system with a high expected number of false signals unnecessarily consumes a large amount of resources.

**Theorem 1** *The optimal solution to the NLP in (3) lies on the boundary of the constraint.*

*Proof.* Define  $S = \{\mathbf{h} : \sum_{i=1}^n \alpha_i \leq \kappa\}$ , and assume that there exists an optimal solution  $P_d(\mathbf{h}^*)$  for some  $\mathbf{h}^* = \{h_1^*, \dots, h_n^*\}$  such that  $\sum_{i=1}^n [1 - F_0(h_i^*)] < \kappa$ . Since  $P_d(\mathbf{h}^*)$  is an optimal solution, then for any other  $\mathbf{h} \in S$

$$\sum_{i=1}^n [1 - F_1(h_i)] p_i \leq P_d(\mathbf{h}^*).$$

But since  $\sum_{i=1}^n \alpha_i < \kappa$  there exists some  $\epsilon > 0$  and some  $\alpha_j$ ,  $j = \{1, 2, \dots, n\}$ , so that  $\alpha'_j = \alpha_j + \epsilon$  and  $\sum_{i \neq j} \alpha_i + \alpha'_j \leq \kappa$ . For any such  $\alpha'_j$  there must exist some  $h'_j = F_0^{-1}(1 - \alpha'_j)$  so that  $1 - F_1(h'_j) \geq 1 - F_1(h_j)$ . Therefore,  $P_d(\mathbf{h}^*)$  is either not the optimal solution or an equivalent solution can be found closer to the boundary. This procedure can be repeated indefinitely until a solution is found for an  $\mathbf{h}$  on the boundary of  $S$ .  $\square$

Now, assuming  $F_0$  is a standard normal distribution and the event of interest manifests itself as a shift in the mean of that distribution, the next lemma shows that the  $n$ -dimensional optimization problem from (3) can be re-expressed as a one-dimensional optimization problem. These assumptions follow from the biosurveillance problem described in Section 2.1.

**Theorem 2** *If  $F_0 = N(0, 1)$  and  $F_1 = N(\gamma, 1)$ ,  $\gamma > 0$ , as in (4), then the optimization problem reduces to finding  $\mu$  to satisfy*

$$\sum_{i=1}^n \Phi\left(\mu - \frac{1}{\gamma} \ln(p_i)\right) = n - \kappa, \quad (5)$$

*and the optimal solution is  $h_i = \mu - \frac{1}{\gamma} \ln(p_i)$ .*

*Proof.* From Theorem 1 we know that the optimal solution lies on the boundary of the constraint, so we can express the constraint in Equation (4) as an equality. The result then follows from reformulating the constrained minimization problem in Equation (4) as the following unconstrained problem:

$$\min_{\mathbf{h}} f = \Phi\left(\Phi^{-1}\left[n - \kappa - \sum_{i=2}^n \Phi(h_i)\right] - \gamma\right) p_1 + \sum_{i=2}^n \Phi(h_i - \gamma) p_i. \quad (6)$$

The partial differential equations with respect to each of the  $h_i$ , for  $i = 2, 3, \dots, n$ , are

$$\frac{\partial f}{\partial h_i} = \frac{\exp\left(-\frac{h_i^2 + \gamma^2}{2}\right) \left(p_i \exp[h_i \gamma] - p_1 \exp\left[\sqrt{2}\gamma \text{Erf}^{-1}\left\{n - 2\kappa - \sum_{i=2}^n \text{Erf}\left[\frac{h_i}{\sqrt{2}}\right]\right\}\right]\right)}{\sqrt{2\pi}}, \quad (7)$$

where  $\text{Erf}(z/\sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) dt$  and  $\text{Erf}^{-1}(\text{Erf}(z)) = z$ .

Now, (7) can be equal to zero only if

$$p_i \exp[h_i \gamma] = p_1 \exp\left(\sqrt{2}\gamma \text{Erf}^{-1}\left\{n - 2\kappa - \sum_{i=2}^n \text{Erf}\left[\frac{h_i}{\sqrt{2}}\right]\right\}\right).$$

Simplifying gives

$$\text{Erf} \left[ \frac{h_i + \frac{1}{\gamma} (\ln(p_i) - \ln(p_1))}{\sqrt{2}} \right] = n - 2\kappa - \sum_{i=2}^n \text{Erf} \left[ \frac{h_i}{\sqrt{2}} \right].$$

Since  $\text{Erf}(z/\sqrt{2}) = 2\Phi(z) - 1$ , after some algebra we have that

$$\Phi \left( h_i + \frac{1}{\gamma} \ln(p_i) - \frac{1}{\gamma} \ln(p_1) \right) + \sum_{i=2}^n \Phi(h_i) = n - \kappa,$$

and substituting  $h_i = \mu - \frac{1}{\gamma} \ln(p_i)$  gives the desired result.  $\square$

One way to think about the one-dimensional optimization in (5) is in terms of finding  $\mu$  such that the sum of the probabilities that each of  $n$  normally distributed random variables (all with the same mean but possibly different variances) is greater than some constant equals  $n - \kappa$ . Specifically, find  $\mu$  such that

$$\sum_{i=1}^n \mathbb{P} \left( X_i > \frac{1}{\gamma} \right) = n - \kappa, \quad (8)$$

where  $X_i \sim N \left( \mu, [\ln(p_i)]^2 \right)$ .

Given the continuity of the normal distribution, (8) makes it clear that an optimal solution is guaranteed to exist. Furthermore, it is a relatively simple problem to solve for  $\mu$  by starting with a large value and gradually decreasing it until the sum of one minus each cdf evaluated at  $1/\gamma$  in (8) equals  $n - \kappa$ .

### 3 Syndromic Surveillance Applications

To illustrate the methodology, in this section we apply it to two hypothetical syndromic surveillance systems.

#### 3.1 Hypothetical Example #1: The 200 Largest US Cities

Based on the assumptions described in Section 2.1, consider a hypothetical syndromic surveillance system for the 200 largest cities in the United States. Assume the probability of attack or outbreak in city  $j$ ,  $p_j$ , is proportional to the population in the city. Of course, the probability of attack could be a function of any number of factors, but for purposes of this example define

$$p_j = \frac{m_j}{\sum_i m_i} = \frac{m_j}{M},$$



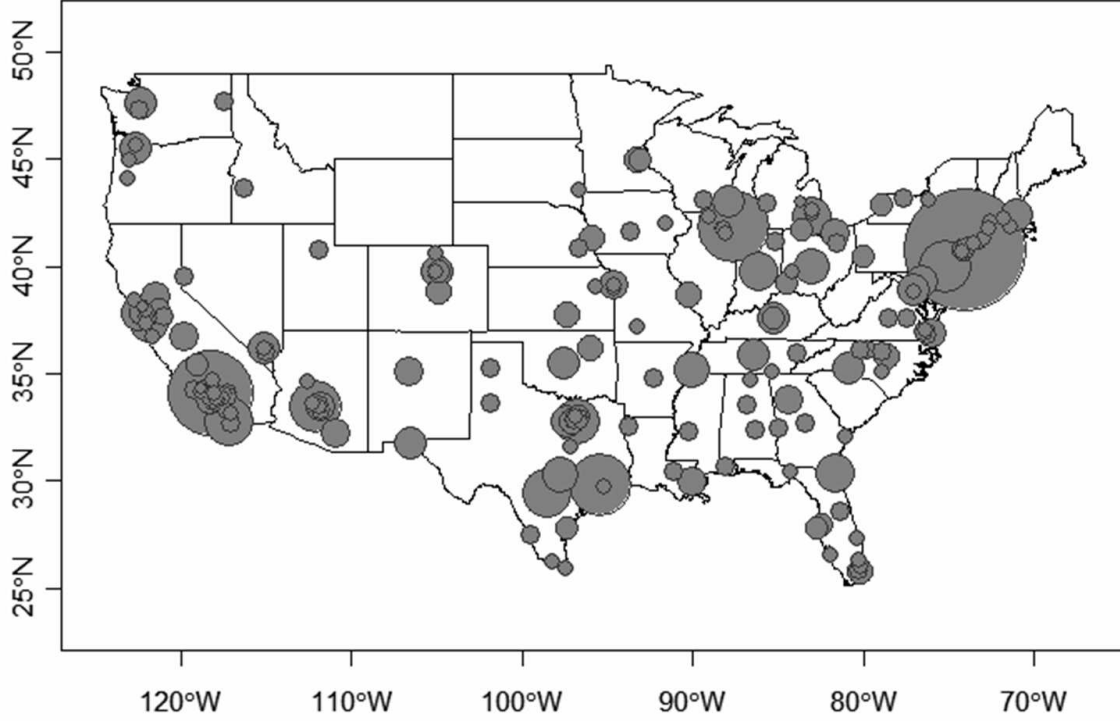


Figure 3: Bubble chart of the 200 largest cities in the United States (Honolulu, Hawaii and Anchorage, Alaska not shown). The bubbles are centered on the cities and their size denotes relative population size.

where  $m_j$  is the population of city  $j$ .

Per the U.S. Census Bureau population estimates for July 1, 2006 ([www.census.gov/popest/cities/SUB-EST2006.html](http://www.census.gov/popest/cities/SUB-EST2006.html)), New York was the largest city with just over 8.2 million people, followed by Los Angeles with just under 4 million, Chicago with just under 3 million, and Houston with just over 2 million. The 200th largest city was West Valley City, Utah with a population of just under 120,000. For a total population of the 200 cities of almost 75 million, our assumption that the probability of attack is simply a function of population size means that the estimated probability of attack for New York is 0.11, Los Angeles is 0.05, Chicago is 0.04, and Houston is 0.03. At the other extreme, West Valley City is 0.002. Figure 3 depicts the data for the 198 cities in the continental United States (Honolulu, Hawaii and Anchorage, Alaska, also in the 200 largest cities, are not shown) using bubbles centered on the cities, where the area of the bubble corresponds to the estimated probability of attack.

Optimizing the system, assuming  $F_0 = N(0, 1)$ ,  $F_1 = N(2, 1)$ , and a maximum expected number of false signals of four per period, the system has a probability of detection of  $P_d = 0.583$ . This is achieved with thresholds ranging from 0.47 for New York, 0.85 for Los Angeles, 1.00 for Chicago, and 1.14 for Houston, to 2.59 for West City Valley, Utah. If one were to have used a common threshold for all the cities of  $h = 2.054$ , which achieves an equivalent expected number of false signals, the probability of detection would decrease 18 percent to  $P_d = 0.478$ . Conversely, setting a common threshold of  $h = 1.79$  to achieve a  $P_d = 0.583$  results in a 59 percent increase in the expected number of false signals to 7.35 per period.

Of course, the choice of four expected false signals per period was made purely for illustrative purposes and assumes that the organization operating the biosurveillance system has the resources and desire to investigate and adjudicate that many signals (on average) per observation period. Figure 4 shows the trade-off between the probability of detection and the expected number of false signals in this scenario. If the organization has additional resources, the constraint on the expected number of false signals can be relaxed and will allow for an increased probability of detection. On the other hand, if the organization is resource constrained, the constraint can be tightened. This will result in a decrease in the probability of detection, but at least all signals will be investigated. *After all, an uninvestigated signal is equivalent to no signal.*

Now, one can easily imagine that operators of a biosurveillance system might want to adjust the system's sensitivity to account for some new intelligence or for other reasons. One way to do this is to adjust  $\mathbf{p}$  to reflect the most recent intelligence about the likelihood of each city being attacked. Another possibility is to introduce additional constraints into the NLP to, for example, ensure that the probability of detection given an attack for some city or cities is sufficiently large.

For example, consider the 200 cities in the previous example, where it is desired that the probability of detection given attack for New York and Washington, D.C. be at least 90 percent. To achieve this requires the addition of two constraints to the NLP in (4):

$$\begin{aligned} h_{NY} &\leq 2 + \Phi^{-1}(0.9) \\ h_{DC} &\leq 2 + \Phi^{-1}(0.9) \end{aligned}$$

The constraints require that the thresholds for New York and Washington be no larger than 0.72. Re-optimizing results in a New York threshold of 0.5 and a Washington threshold of 0.72. For the other cities, new thresholds ranged from 0.87 for Los Angeles, 1.03 for Chicago, and 1.17 for Houston, to 2.61 for West City Valley, Utah. The overall probability of detection decreases slightly

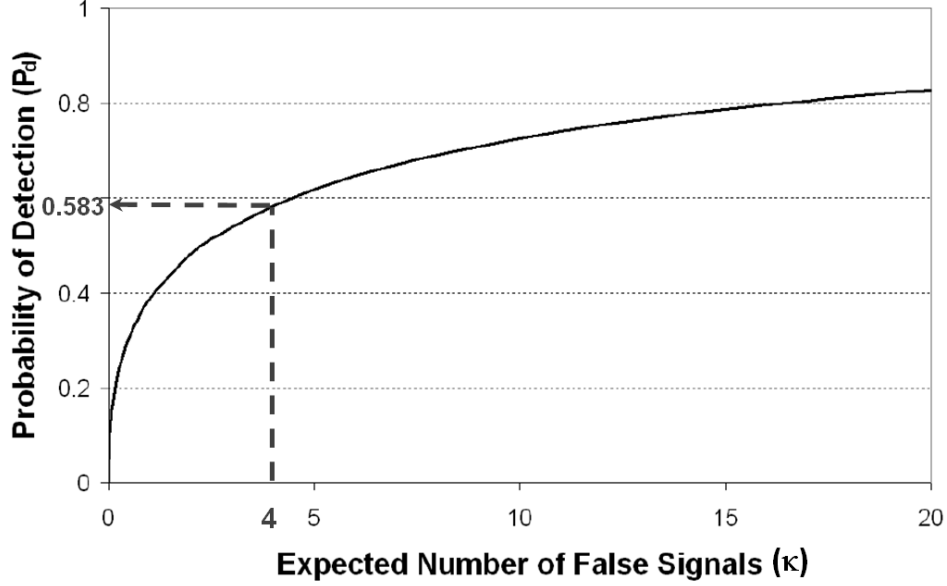


Figure 4: The trade-off between the expected number of false signals and the probability of detection for the optimal thresholds for Example #1. For  $\kappa = 4$  the optimal thresholds give  $P_d = 0.583$ . Increasing  $\kappa$  increases the probability of detection, but with decreasing returns.

to  $P_d = 0.578$ .

### 3.2 Hypothetical Example #2: Monitoring 3,141 US Counties

In Example #1, one might take exception to only monitoring the 200 largest cities. The implicit assumption is that there is zero probability of an attack outside of these cities. One alternative would be to field a biosurveillance system designed to monitor all 3,141 counties in the United States. For the purposes of illustration, as with Example #1, we use the proportion of the total population in a county as a surrogate for the probability that county is attacked.

Per the U.S. Census Bureau county population estimates for 2006 ([www.census.gov/popest/counties/files/CO-EST2006-ALLDATA.csv](http://www.census.gov/popest/counties/files/CO-EST2006-ALLDATA.csv), “poestimate2006”), Los Angeles was the largest county with just under 10 million people, followed by Cook county with just under 5.3 million, and Harris county with just under 4 million. The smallest county was Loving county, Texas with a population of 60. For a total United States population in 2006 of 299.4 million, the estimated probability of attack ranges from Los Angeles county at 3.3 percent to Loving county at 4 one hundred thousandths of a percent.

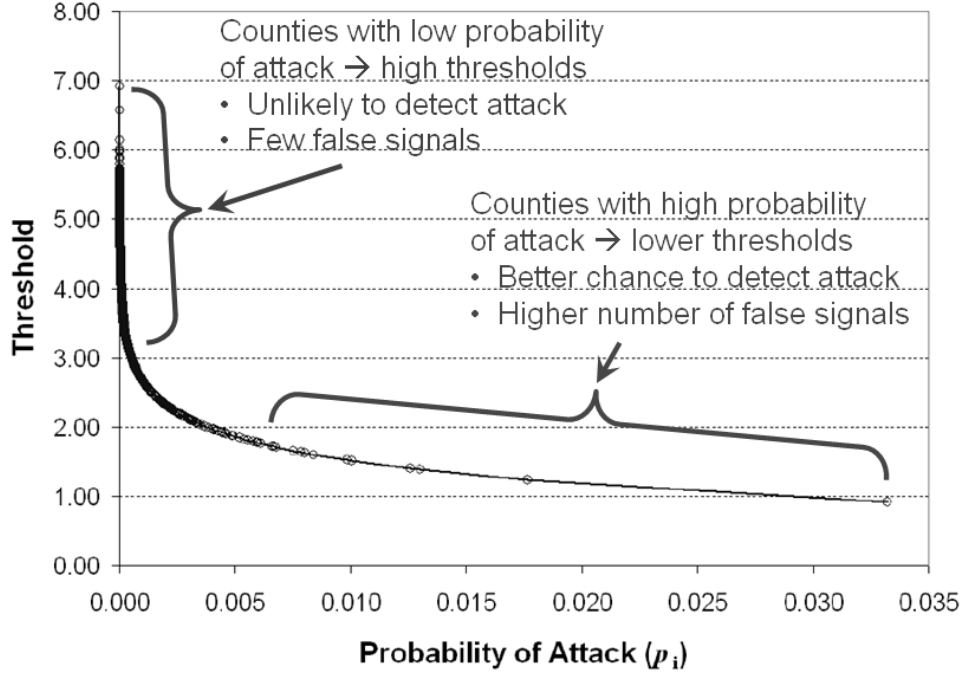


Figure 5: Plot of the optimal thresholds versus probability of attack for Example #2. The optimized thresholds focus surveillance on those locations with higher probability of attack.

If we assume as before that  $F_0 = N(0, 1)$ ,  $F_1 = N(2, 1)$ , and a maximum on the expected number of false signals of four per period, the system has a probability of detection of  $P_d = 0.333$ . This is achieved with thresholds ranging from 0.91 for Los Angeles county, 1.23 for Cook county, and 1.38 for Harris county, to 6.92 for Loving county. Figure 5 shows a plot of the optimal thresholds versus the probability of attack and Figure 6 is a map showing the probability of attack and thresholds by county.

The cost for increasing the number of regions being monitored from 200 to 3,141 is about a 43 percent (25 percentage point) decrease in the probability of detecting an attack that manifests itself as a two standard deviation increase in the mean of the residuals. The benefit is an increase in the area being monitored. Of course, this is something of an apples-to-oranges comparison since in the 200-cities example the probability of detecting an attack is conditional on the attack occurring within that region. Thus, there are large areas of the country for which an attack could not be detected at all. In contrast, the county-level system has some power to detect an attack anywhere in the United States, but this comes at the expense of the power to detect an attack within the 200-cities region.

In terms of the county-level model, it is worth noting that while those counties with very low probabilities of attack have such high thresholds that they will be virtually unable to detect a moderately-sized outbreak/attack, these counties are being monitored at a level consistent with their risk of attack. That is, the optimization has made the necessary trade-off of probability of detection versus the likelihood of false signals in order to maximize the probability of detecting an attack somewhere in the country within a manageable false signal rate.

Now, consider system performance if one were to have used a common threshold for all the counties of  $h = 3.018$ , which achieves the same expected number of false signals (four per period), the probability of detection would be cut more than in half to  $P_d = 0.154$ . This decrease in sensitivity occurs because the system is less able to detect an attack in those locations most likely to be attacked. Conversely, setting a common threshold of  $h = 2.433$  to achieve a  $P_d = 0.333$  results in an almost six-fold increase in the expected number of false signals to 23.5 per period.

### 3.3 Discussion

In Examples #1 and #2 the thresholds were set assuming  $\kappa = 4$  and  $\gamma = 2$ . Choosing  $\kappa$  is a matter of resources and should be based on an organizational assessment of the average number of signals that can be investigated per period. For a fixed number and type of sensors, one can improve the system-wide probability of detection by increasing the expected number of false signals allowed. As shown in Figure 4, however, there is a decreasing level of improvement in the probability of detection for resources invested in adjudicating signals. Table 2 shows the trade-off in probability of detection for the 200-cities example for four levels of  $\gamma$  and for five values of  $\kappa$ .

Choosing the value of  $\gamma$  over which to optimize is a subjective judgement based on the likely magnitude of the mean increase that is important to detect. As shown in Table 2, once the choice is made and the thresholds set, an outbreak manifested as a small value for  $\gamma$  (relative to the standard deviation of the observations or residuals) will be harder to detect and will result in a lower probability of detection. Conversely, an outbreak manifested as a larger  $\gamma$  will make it easier to distinguish between  $F_0$  and  $F_1$  and thus will result in a higher the probability of detection.

That said, a relevant question is how sensitive the resulting probability of detection is to the mis-specification of  $\gamma$  during the optimization. For example, what happens if the thresholds are chosen using an optimization based on  $\gamma = 2$  and then the actual outbreak manifests itself with  $\gamma = 1$  or  $\gamma = 3$ ? Table 3 shows the actual probabilities of detection that would occur for the 200-cities example using the optimal thresholds determined for  $\gamma = 2$ . Comparing Table 3 to Table 2 we see

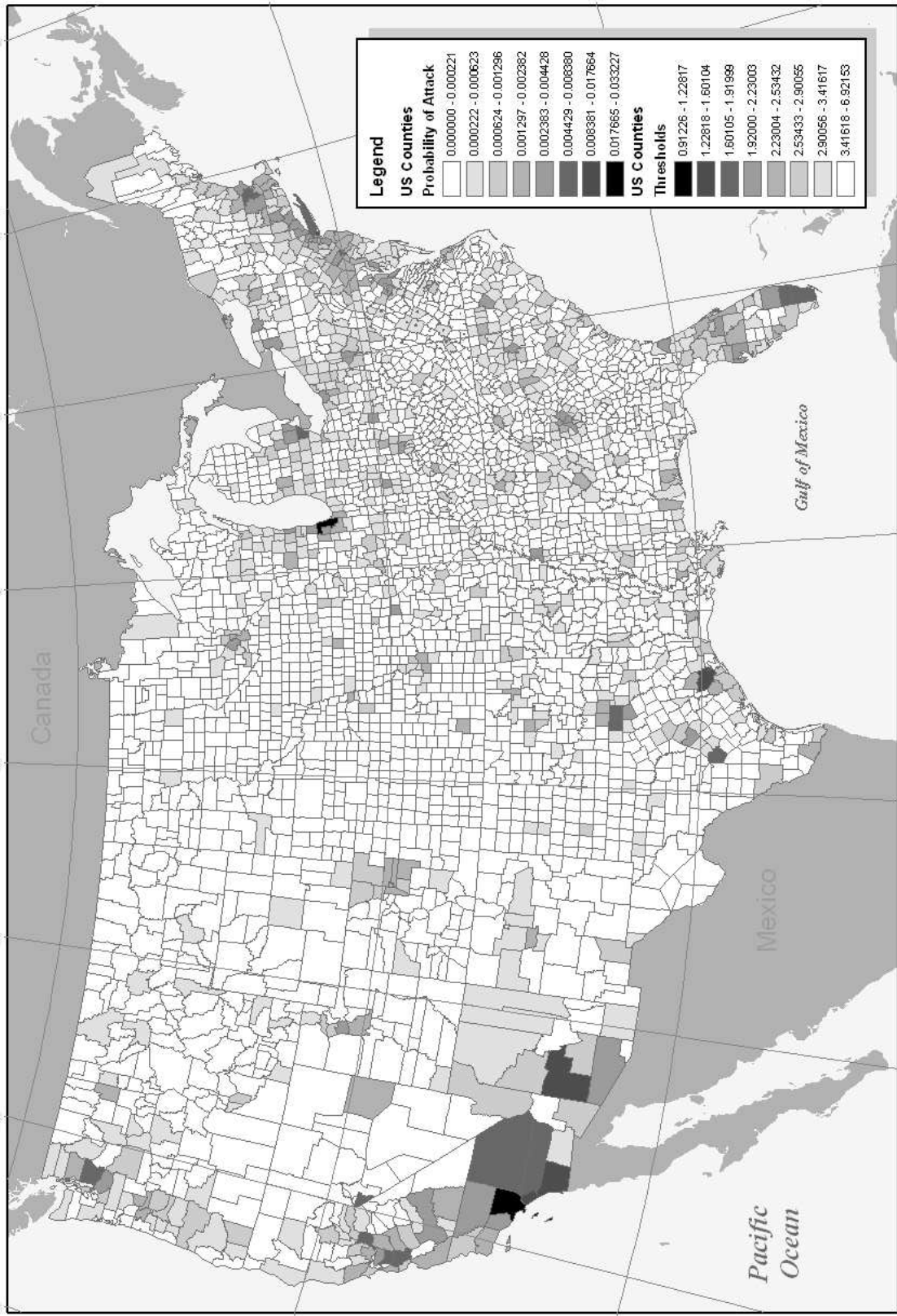


Figure 6: Map of the optimal thresholds and associated probabilities of attack from Example #2 for the counties in the contiguous continental United States.

Table 2: Optimal probabilities of detection in the 200-cities example for various values of  $\gamma$  and  $\kappa$ .

$P_d$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
$\gamma = 1$	0.165	0.228	0.272	0.307	0.336
$\gamma = 2$	0.388	0.481	0.540	0.583	0.618
$\gamma = 3$	0.726	0.801	0.840	0.866	0.885
$\gamma = 4$	0.939	0.964	0.974	0.980	0.984

Table 3: Actual probabilities of detection in the 200-cities example when the system is optimized for  $\gamma = 2$  and the outbreak/attack results in  $F_1$  with  $\gamma$  as shown in the left column of the table.

$P_d$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
Observed $\gamma = 1$	0.137	0.193	0.235	0.269	0.298
Observed $\gamma = 2$	0.388	0.481	0.540	0.583	0.618
Observed $\gamma = 3$	0.711	0.790	0.832	0.859	0.879
Observed $\gamma = 4$	0.925	0.955	0.968	0.976	0.981

that there is some degradation in  $P_d$  if the actual outbreak manifests at some  $\gamma$  other than the value used to optimize the system, but the loss in detection probability is not large.

For biosurveillance system designers and operators, it is important to understand the interplay between probability of detection and the expected number of false signals. In Figure 4 we have already seen that, after a certain level, improving the probability of detection requires an increasingly larger expected number of false signals. A similar result holds when one tries to decrease the thresholds in order to achieve higher probabilities of detection. For example, Figure 7 demonstrates how the probability of detection and expected number of false signals change when the optimal thresholds from Example #1 are uniformly lowered by the percentages indicated on the horizontal axis. In the plot, zero percent decrease corresponds to the probability of detection and expected number of false signals for the optimal thresholds and a 100 percent decrease in thresholds gives the probability of detection and expected number of false signals when all the thresholds are set to 0. What we see is that the expected number of false signals rises much faster than the probability of detection for threshold decreases of more than 20 percent or so.

Similarly, if a system's thresholds are not carefully set and controlled then it is possible for the number of false signals to rapidly exceed the available resources to adjudicate them. To illustrate this, we conducted a simple simulation in which the optimal thresholds from Example #1 were

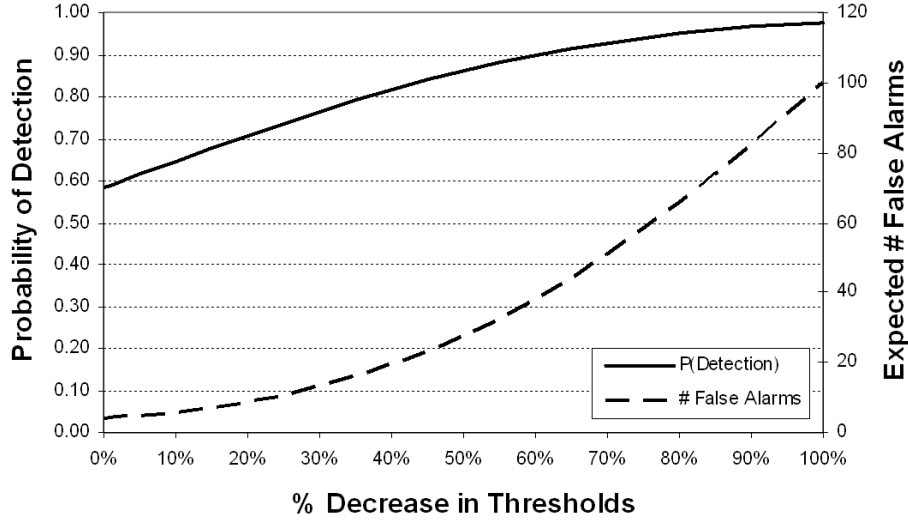


Figure 7: Changes in the probability of attack and expected number of false signals for Example #1 when the optimal thresholds are uniformly decreased by some percentage as shown on the horizontal axis.

randomly varied by a certain percentage. Figure 8 shows that when the thresholds are allowed to randomly vary anywhere from five to 200 percent of their optimal values, the *average* system-wide probability of detection is essentially unaffected. However, Figure 8 also shows that as the fluctuation increases the expected number of false signals increases significantly. In fact, allowing the optimal thresholds to vary randomly by 200 percent raises the average number of false signals by nearly sixteen hundred percent, from four expected false signals to sixty-two. It thus behooves biosurveillance system architects to both carefully design and control the system in order to manage the number of false signals the system will generate.

Finally, we also explored how a biosurveillance system might perform if the thresholds were calculated assuming the standardized residuals were normally distributed but the actual distribution violated that assumption. In particular, using the 200-cities example we allowed the standardized residuals to follow a  $t$ -distribution with various degrees of freedom and then compared system performance with the thresholds appropriately optimized for the  $t$ -distribution to thresholds set using Theorem 2 assuming the residuals were normally distributed.

Table 4 shows the results, where the expected number of false signals were constrained to one per period (i.e.,  $\kappa = 1$ ) and we set  $\gamma = 2$ . The first column labeled “df” gives the degrees of freedom for the  $t$ -distribution, which we varied from  $\infty$  (i.e., a standard normal) to  $\text{df}=1$ . The next two



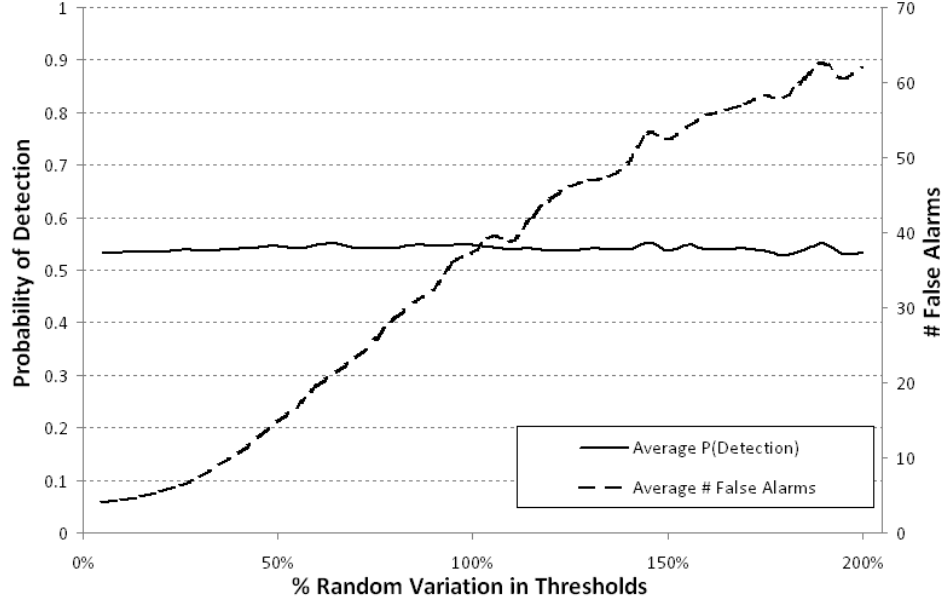


Figure 8: Effect on the probability of detection and expected number of false signals in Example #1 when individual thresholds are allowed to randomly vary from five percent to 200 percent.

columns give the system performance, in terms of  $P_d$  and  $\kappa$ , for the optimal thresholds calculated for the correct  $t$ -distribution (where we used the Excel Solver, as described in Section 2.2.1). Here we see, not surprisingly, that  $P_d$  decreases for decreasing degrees of freedom (and fixed  $\kappa$ ), since decreasing degrees of freedom corresponds to heavier tails and thus more variability.

In the last two columns of Table 4 we see how the system would perform if the thresholds were set using the Theorem 2; i.e., incorrectly assuming the residuals followed a standard normal distribution. What is most interesting is that  $P_d$  changes very little while the observed average number of false signals significantly increases as the distribution is increasingly misspecified. Compared to the optimal  $P_d$ s, using the incorrect thresholds results in higher  $P_d$ s, the cost of which at the most extreme (i.e.,  $df=1$ ) is a 22-fold increase in the average number of false signals over what was desired.

Table 4 reinforces what we already observed in Figure 7: the false signal rate is much more sensitive to the choice of thresholds than is the probability of detection. Said another way, biosurveillance system designers and operators should be very cautious about how thresholds are chosen since small changes that have minimal effect on detection performance can have large effects on the number

Table 4: Performance of a system in the 200-cities example (with  $\gamma = 2$  and  $\kappa = 1$ ) when it is optimized for residuals that follow a  $t$ -distribution with  $df$  degrees of freedom compared to its performance when using thresholds calculated assuming the residuals follow a standard normal distribution.

Results for Correct Thresholds for $t$ -distribution			Results for Incorrect Thresholds Based on Normal	
df	$P_d$	$\kappa$	$P_d$	Observed $\kappa$
$\infty$	0.388	1.000	0.388	1.000
500	0.385	1.000	0.388	1.021
50	0.363	1.000	0.388	1.221
25	0.340	1.000	0.389	1.471
10	0.290	1.000	0.391	2.380
5	0.247	1.000	0.395	4.279
2	0.199	1.000	0.404	11.136
1	0.173	1.000	0.416	22.136

of false signals a system produces. In addition, for those using the EARS' C1 and C2 algorithms (including the W2 in BioSense), this example suggests caution in using the results of Theorem 2 to set thresholds. For these algorithms, because the denominator in the statistics is an estimator for the residual standard deviation based on seven observations, the statistics being tracked may be more likely to follow a  $t$ -distribution than a normal distribution.

For additional discussion, examples, and more detail on the application of this methodology to biosurveillance, see Banschbach (2008).

## 4 Summary and Conclusions

In this paper we have described a framework for optimizing thresholds for a system of biosurveillance or other threshold detection sensors. In so doing, we have made a number of assumptions about the sensor system, including that we can appropriately model and remove any systematic effects in the data from  $n$  sensors so that the resulting residuals are independent and that the sensor signals are independent over time. We have also assumed identical distributions across all of the sensors and, in most of our examples, that these are normal distributions with the event of interest manifesting itself as an increase in the mean of the  $F_0$  distribution.

The choice of the normal distribution was driven by previous work which showed that monitoring the residuals from an adaptive regression could be an effective biosurveillance strategy, and it is not unreasonable to assume such residuals follow (at least to a first approximation) a normal distribution. However, the methodology described herein is not limited to this assumption, nor does it require identical distributions for all of the sensors. What is required is that the probability of exceeding a given threshold can be calculated for each sensor when no event being present (a false signal) and when an event of interest is present (a true signal).

The assumption that sensor signals are independent over time simplified the optimization calculations and may or may not reflect real-world conditions for a given biosurveillance or other sensor system. Our motivation was biosurveillance in which some of the algorithms currently in use are of this type. However, there are other methods that use both current and historical information (such as the CUSUM and EWMA quality control methods) for which additional research is required to determine how to implement an equivalent approach. Certainly the idea is relevant—those methods also use thresholds to reach a binary decision—but because the distribution at each time period is conditional on the history up to that time period, no simple expressions for the percentiles and probabilities exist.

In some sensor systems it may be by design to have multiple sensors in the same location all monitoring for an event of interest in that region. In this situation, it is quite likely—even desirable—that the sensors’ signals are correlated. In these systems, the signals from the various sensors are fed into some sort of “fusion center” from which a single determination is made about whether an event of interest has occurred in the region. In such systems, it would be inappropriate to use the methodology described herein to develop thresholds for the individual sensors. Rather, if the fusion center’s output is based on a threshold detection methodology of the combined sensor inputs, then this methodology should be used to optimize the fusion center thresholds.

In terms of the biosurveillance problem, note that in a real surveillance system each hospital will be monitoring  $m$  different syndromes simultaneously. Thus, if the total number of system wide false signals that can be tolerated per period is  $\kappa$ , the thresholds for each syndrome must be optimized subject to  $\kappa/m$  expected number of false signals. Of course, this assumes that it is equally important to detect an anomaly in one syndrome as in any other syndrome. If this is not the case, it is also possible to set the allowable expected number of false signals differentially by syndrome, where the higher the number of false signals allowed, the more sensitive the overall system will be to detecting a true outbreak of that particular syndrome.

We conclude by stressing that this methodology does not apply just to biosurveillance systems. Systems of sensors have historically been used in military applications and today, with increasing computing power and miniaturization, the uses of systems of sensors are proliferating well beyond the military. Examples include such diverse applications as meteorology, supply chain management, equipment and production monitoring, health care, production automation, traffic control, habitat monitoring, and health surveillance. See, for example, Gehrke & Liu (2007), Xu (2007), Intel (2007), Trigoni (2004), and Bonnet (2004). This methodology can potentially be applied to any such application that uses threshold detection-based sensors.

This methodology also has promise in industrial quality control for optimizing Shewhart chart applications. Consider, for example, a factory with  $n$  production lines, each monitored by a single Shewhart chart, where for whatever reason one of the lines is more likely to go “out of control” compared to the others. Using standard practices, the factory would probably set the thresholds equally on all the Shewhart charts. However, that would mean less-than-optimal factory performance since ideally one would want to tune the control limits to be more sensitive to catching the line more likely to go out-of-control. The methodology presented in this paper provides the means for optimizing the thresholds. It would require a change in the way one thinks about the design of control charts since the objective function and constraint are not in the usual terms of in-control and out-of-control average run lengths. In addition, one would need to develop a methodology for estimating the probability that each line goes out of control. However, these are subjects for another paper.

## Acknowledgments

We thank Matt Carlyle and Johannes Royset for their insights into this problem and their comments on an earlier version of the paper. We also thank three anonymous reviewers for their helpful comments that significantly improved the paper. R. Fricker’s work on this effort was partially supported by Office of Naval Research grant N0001407WR20172.

## References

- Banschbach, D.C. 2008. *Optimizing Systems of Threshold Detection-Based Sensors*. Monterey, CA: Master’s Thesis, Naval Postgraduate School.
- Bonnet, P. 2004. *Sensor Network Applications*. Accessed on-line at [www.diku.dk/undervisning/2004v/336/Slides/SN-applications.pdf](http://www.diku.dk/undervisning/2004v/336/Slides/SN-applications.pdf) on October 2, 2007.

- CDC. 2003. *Syndrome Definitions for Diseases Associated with Critical Bioterrorism-associated Agents*. Accessed on-line at [www.bt.cdc.gov/surveillance/syndromedef/](http://www.bt.cdc.gov/surveillance/syndromedef/) on November 21, 2006.
- CDC. 2004. Syndromic Surveillance: Reports from a National Conference, 2003. *Morbidity and Mortality Weekly Report*, **53**(supplement).
- CDC. 2006a. *BioSense System Website*. Accessed on-line at [www.cdc.gov/biosense/](http://www.cdc.gov/biosense/) on April 30, 2007.
- CDC. 2006b. *Early Aberration Reporting System Website*. Accessed on-line at [www.bt.cdc.gov/surveillance/ears/](http://www.bt.cdc.gov/surveillance/ears/) on November 22, 2006.
- CDC. 2008. *BioSense System Website*. Accessed on-line at [www.cdc.gov/biosense/publichealth.htm](http://www.cdc.gov/biosense/publichealth.htm) on October 8, 2008.
- Chair, Z., & Varshney, P.K. 1986. Optimal Data Fusion in Multiple Sensor Detection Systems. *IEEE Transactions on Aerospace and Electronic Systems*, **AES-22**(1), 98–101.
- Fricker, Jr., R.D. 2007. Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance. *Advances in Disease Surveillance*, **3**. Available on-line at [www.isdsjournal.org](http://www.isdsjournal.org).
- Fricker, Jr., R.D. 2008. Syndromic Surveillance. *Encyclopedia of Quantitative Risk Assessment*, 1743–1752.
- Fricker, Jr., R.D., & Rolka, H.A. 2006. Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance. *Chance*, **91**, 4–13.
- Fricker, Jr., R.D., Hegler, Benjamin L., & Dunfee, David A. 2008a. Comparing Biosurveillance Detection Methods: EARS' Versus a CUSUM-based Methodology. *Statistics in Medicine*, **27**, 3407–3429.
- Fricker, Jr., R.D., Knitt, Matthew C., & Hu, Cecilia X. 2008b. Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance. *Quality Engineering*, **20**, 478–494.
- Gehrke, J., & Liu, L. 2007. *Sensor Network Applications*. Accessed on-line at [http://dsonline.computer.org/portal/site/dsonline/menuitem.6dd2a408dbe4a94be487e0606bcd45f3/index.jsp?&pName=dso\\_level1\\_article&TheCat=1015&path=dsonline/2006/04&file=w2gei.xml&](http://dsonline.computer.org/portal/site/dsonline/menuitem.6dd2a408dbe4a94be487e0606bcd45f3/index.jsp?&pName=dso_level1_article&TheCat=1015&path=dsonline/2006/04&file=w2gei.xml&) on October 2, 2007.
- Hutwagner, L., Thompson, W., Seeman, G.M., & Treadwell, T. 2003. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **80**(No. 2, Supplement 1), 89i–96i.
- Intel. 2007. *Sensor Nets/RFID Website*. Accessed on-line at [www.intel.com/research/exploratory/wireless\\_sensors.htm](http://www.intel.com/research/exploratory/wireless_sensors.htm) on October 2, 2007.
- International Foundation For Functional Gastrointestinal Disorders. 2006. Accessed on-line at [www.iffgd.org/GIDisorders/glossary.html](http://www.iffgd.org/GIDisorders/glossary.html) on November 21, 2006.
- Kress, M., Szechtman, R., & Jones, J.S. 2008. Efficient Employment of Non-Reactive Sensors. *Military Operations Research*, **13**(4), 45–57.
- Montgomery, D.C. 2001. *Introduction to Statistical Quality Control*. Fourth edn. New York, NY: John Wiley & Sons.
- Shewhart, W.A. 1931. *Economic Control of Quality of Manufactured Product*. Princeton, NJ: D. van Nostrand Company, Inc.
- Shmueli, Galit. 2006. Accessed on-line at <https://wiki.cirg.washington.edu/pub/bin/view/Isds/SurveillanceSystemsInPractice> on October 8, 2008.
- Tokars, J. 2006a. Personal communication, December 28, 2006.
- Tokars, J. 2006b. The BioSense Application. PHIN Conference. Accessed on-line at <http://0-www.cdc.gov.mill1.sjlibrary.org/biosense/files/Jerry.Tokars.ppt#387,1,TheBioSenseApplication> on November 27, 2006.
- Toprani, A., Ratard, R., Straif-Bourgeois, S., Sokol, T., Averhoff, F., Brady, J., Staten, D., Sullivan, M., Brooks, J.T., Rowe, A.K., Johnson, K., Vranken, P., & Sergienko, E. 2006. Surveillance in Hurricane

- Evacuation Centers - Louisiana, September-October 2005. *Morbidity and Morality Weekly Report*, **55**, 32-35.
- Trigoni, N. 2004. *Sensor Networks: Applications and Research Challenges*. Accessed on-line at <http://locationprivacy.objectis.net/talks/5trigoni> on October 2, 2007.
- Washburn, A.R. 2002. *Search and Detection*. Fourth edn. Topics in Operations Research Series. Linthicum, MD: INFORMS.
- Xu, N. 2007. A Survey of Sensor Network Applications. Accessed on-line at <http://courses.cs.tamu.edu/rabi/cpsc617/resources/sensor%20nw-survey.pdf> on October 2, 2007.
- Zhu, M., Ding, S., Brooks, R.R., Wu, Q., Rao, N.S.V., & Sitharama Iyengar, S. 2007. Fusion of Threshold Rules for Target Detection in Sensor Networks. Accessed on-line at [www.csc.lsu.edu/~iyengar/final-papers/FusionTOSN.pdf](http://www.csc.lsu.edu/~iyengar/final-papers/FusionTOSN.pdf) on October 2, 2007.